

Weekly Report

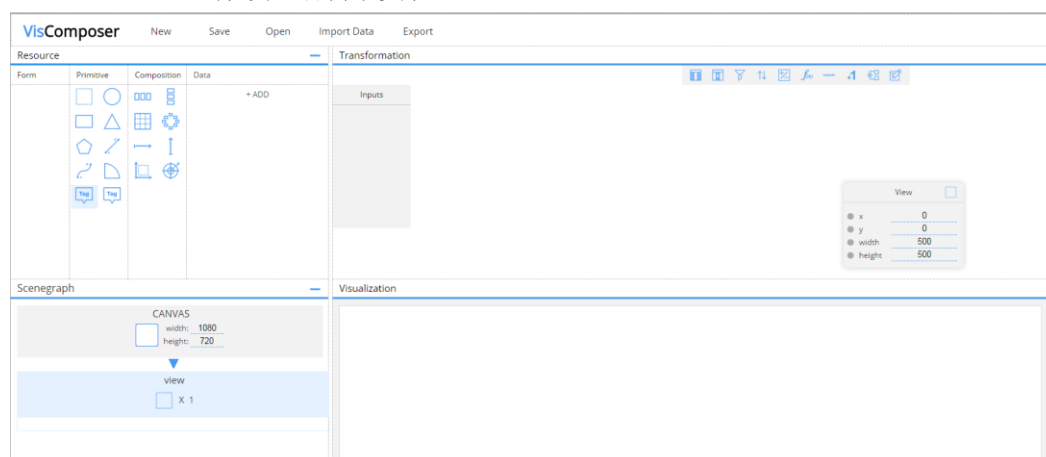
胡万祺

一、 本周工作

【VisComposer】

1. 写专利说明书，以下是列出的四个步骤，后面主要是围绕这四个步骤写。
 - a) 将信息可视化设计流程规范化，分成了数据读取、数据变换、视觉映射和视图绘制四个阶段，每个信息可视化的设计流程都要包含这四个阶段。每个阶段相互独立，前一个阶段的输出作为后一个阶段的输入。
 - b) 为上述流程的构建提供交互界面，自带解析读取数据文件、图形绘制的功能，并预定义多种不同功能的数据变换、图元（视觉映射）模块供自由选择。构建流程采取流程图连线的方式，通过交互直接指定各数据变换、图元模块输入输出，自由构建可视化设计流程并生成可视化视图。
 - c) 预定义的数据变换、图元模块开放代码编辑接口，可用 javascript 代码自由修改和增加模块功能；另有自定义模块，用 javascript 代码指定输入输出，并实现中间处理过程。
 - d) 自定义设计流程保存成模板，以便进行迭代设计。写得比较慢，现在还未做润色和配图。预计下周中写完。

2. 目前 viscomposer 重构版系统已经跑通场景树初始化->添加空白视图节点->生成 workflow。还有最后绘制未实现。



【农业大数据】

1. 数据整理及自动化处理（陈俊）

根据新加坡那边郭奇博士的要求，将之前爬取的数据整理成了符合要求的形式（每个城市或市场的每个产品一个文件），主要是先将数据入库，然后再利用数据库接口重新整理。

目前针对一亩田和农产品和批发市场价格信息网，在 node4 节点机上实现了每天定时爬取并且入库的自动化流程。

```
# m h dom mon dow   command
0 22 * * * /home/zhang/just-cj/WebCrawler/jgsb/run_jgsb
0 21 * * * /home/zhang/just-cj/WebCrawler/ymt/run_ymt
```

为了保证当天数据尽可能完整的被爬取下来，设置这两个网站的任务在晚上 9 点和 10 点执行。run_jgsb 和 run_ymt 是两个网站对应任务的 shell 脚本。

以其中一个为例，该脚本主要完成了爬取和入库的操作。

```
1 #!/bin/bash
2 cd /home/zhang/just-cj/WebCrawler/ymt/code/
3 node test.js > ymt.log
4 cd /home/zhang/just-cj/WebCrawler/ymt/toDB
5 node toDB.js > toDB.log
6
```

此外，重新整理数据文件并没有每天在做，主要原因是现在的数据量比较少，建议积累一定程度的数据以后在整理数据比较好。

2. 需求说明书分析（陈俊）

研究了之前的需求说明书，整理了一些需要补充的地方，如下：

- 补充用例图及用例
为每个系统角色补充用例图以及用例说明
- 增加输入输出要求，并补充数据流程图
补充以下内容
 - 数据流定义
 - 数据元素定义
 - 附录中添加数据流程图
- 补充状态图
可以直观地了解系统的运作。
- 完善性能需求部分
这一部分应当更为具体，包括对于数据精度以及响应时间的具体要求。

3. 分布式存储系统实现（王艺）

在上次做完分布式后插入了 8000w+ 数据量做测试。Mongos 会调节所有 shard，使每个 shard 有相同的 chunk 数目，而不是平衡分片的大小。

mongodb 分片的条件是片键，片键拥有更高的基数时，分片也会越均匀。最近是用时间作为片键，但问题是，时间的基数虽然很大，但是仍然会有 chunk 太大无法移动的报错。

在目前的集群中分布如下：

```

shard set10 at set10/VAG-Node10:27002,VAG-Node11:27001,VAG-Node12:27000
data : 2.05GiB docs : 19686068 chunks : 45
estimated data per chunk : 46.72MiB
estimated docs per chunk : 437468

shard set11 at set11/VAG-Node10:27000,VAG-Node11:27002,VAG-Node12:27001
data : 2.22GiB docs : 21375655 chunks : 57
estimated data per chunk : 40.05MiB
estimated docs per chunk : 375011

shard set12 at set12/VAG-Node10:27001,VAG-Node11:27000,VAG-Node12:27002
data : 1023.7MiB docs : 9584265 chunks : 44
estimated data per chunk : 23.26MiB
estimated docs per chunk : 217824

shard set7 at set7/VAG-Node7:27002,VAG-Node8:27001,VAG-Node9:27000
data : 1.65GiB docs : 15872763 chunks : 44
estimated data per chunk : 38.53MiB
estimated docs per chunk : 360744

shard set8 at set8/VAG-Node7:27000,VAG-Node8:27002,VAG-Node9:27001
data : 974.01MiB docs : 9119007 chunks : 44
estimated data per chunk : 22.13MiB
estimated docs per chunk : 207250

shard set9 at set9/VAG-Node7:27001,VAG-Node8:27000,VAG-Node9:27002
data : 881.2MiB docs : 8250128 chunks : 44
estimated data per chunk : 20.02MiB
estimated docs per chunk : 187502

Totals
data : 8.75GiB docs : 83887886 chunks : 278
Shard set10 contains 23.46% data, 23.46% docs in cluster, avg obj size on shard : 112B
Shard set11 contains 25.48% data, 25.48% docs in cluster, avg obj size on shard : 112B
Shard set12 contains 11.42% data, 11.42% docs in cluster, avg obj size on shard : 112B
Shard set7 contains 18.92% data, 18.92% docs in cluster, avg obj size on shard : 112B
Shard set8 contains 10.87% data, 10.87% docs in cluster, avg obj size on shard : 112B
Shard set9 contains 9.83% data, 9.83% docs in cluster, avg obj size on shard : 112B

```

数据库在空闲时会自动进行均衡，即调整 chunk 的位置，sh.status() 可以看到运行状态，目前主要报错是两种，一种是 chunk too big to move，还有一种是目标分片正在删除 chunk 无法接受其他分片转移的 chunk。目前的数据放在集群里已经两天半，还是在均衡中，chunk 移动很缓慢。

4. js convention 文档（王艺）